wrangle_report

July 29, 2019

1 Wrangle Report

1.0.1 By Diego ZUNIGA

During this project a database is built gathering data from different sources. Then the data is assesses and finally cleaned to produce some insights and visualizations. Now the followed process will be described.

1.1 Gathering

In order to get all the information necessary to build the database, the information was collected from three different sources. The first one is a file with the tweeter information from the account WeRateDogs, the second one is a file coming from some automatic classification algorithms and the third one was actually the result of querying the tweeter API. In order to query the API of tweeter it was necessary to create an account for developers and get permissions to send some requests. After that using the python's library Tweepy the information of all the tweets was consulted using their ids.

1.2 Assesing

After gathering the data it was assessed by visual inspection, and also programatically using some functions included in the pandas library. This assesing proccess gave as result some issues that are listed below. But it is important to notice that only some issues were documented and it doesn't represent the totallity of the necessary for completely wrangling the data.

1.2.1 Quality

Table 1

- Denominator not normalized
- 'retweeted_status_user_id' and 'retweeted_status_id' are in scientific notation, it should be a string
- Some names are None and there are many 'a' and 'the' names which is unusual.
- tweet_id is an integer
- 'in_reply_to_status_id', 'in_reply_to_user_id', 'retweeted_status_id', 'retweeted_status_user_id' and 'retweeted_status_timestamp' have too few data to be relevant.
- the minimum value in denominator is 0 which is not possible.

Table 2

- tweet_id is an integer

Table 3

- id column should be called tweet_id
- id should be string instead of integer
- lang should be called language

1.2.2 Tidyness

Table 1

- Type of dog should be one column
- Rating numerator and denominator can become one column with a normalized value

Table 2

- This table can be merged into table 1

Table 3

- This table can be merged into table 1
- There are retweets and we don't want them

All those issues where then solved in the cleaning process.

1.3 Cleaning

During the cleaning proccess, the found issues were not solved in the order they are presented, because sometimes while solving a tidyness problem some other quality issues disappeared. But in any case all of them where mentioned and solved. In general every problem is taken one by one, then a solution is proposed in the definition, after the code is written in order to solve the problem itself and finally the data is assessed to see if the code did what it was intended to do.

1.4 Storing

Once the data is cleaned and ordered in one dataframe, it is stored in a file in format .csv which is supposed to be a clean and well wrangled dataframe. But as was already said, it is not exhaustive and it doesn't cover all the problems that the data can contain.

1.5 Insights

After the data is wrangled it is possible to analyse it and produce some insights and visualisations. So with this purpose some variables were compared and it was found that the pupper dog type is the most common. And also the twitter rating of the dogs was compared with the tweeter measures like favorite counter and retweet counter finding a not very strong tendency but still some. It shows that when dogs were well rated by the owners who posted the tweets, it is more likely to find higher values of retweet counters and the same for the favorite counters. Which might mean that well rated dogs by their owners are more likely to be well appreciated by the social network community.